Bayesian Network Modeling of Signaling Pathways and High-Throughput Data

Peter Woolf

University of Michigan

Abhik Shah
Mirka Utzka
Angel Alvarez
Raghu Kainkaryam
Weijun Luo

Andrew Hodges
Yongquin He
Zuoshuang Xiang
Alfred Hero
Barry Taylor

Dongxiao Zhu
Bijun Tan
Jignesh Patel
Sandeep Tata
Ryan Welch

Team Goals

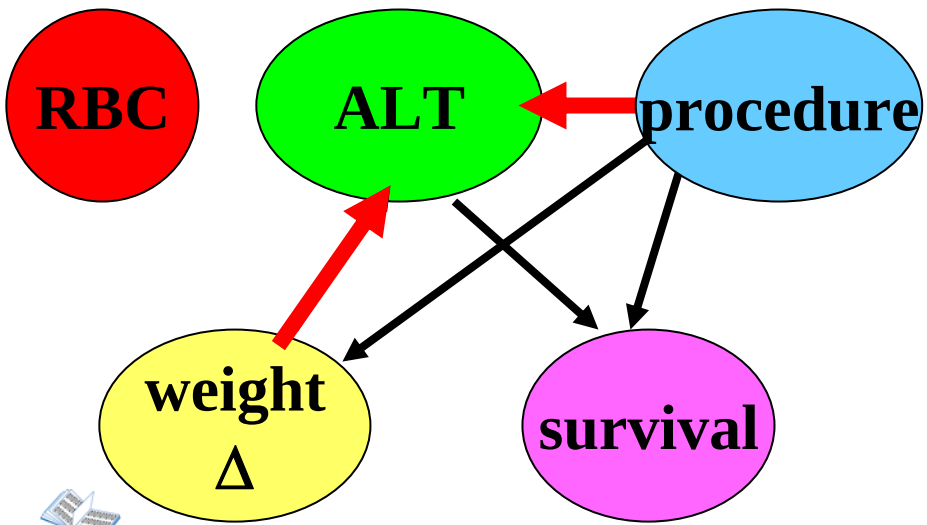
- Provide a flexible, data driven bioinformatic modeling environment that assists users in:
  - Integrating diverse sources of data
  - Interpreting large data sets
  - Making quantitative predictions

Key modeling tool: Bayesian Networks

Bayesian Networks: General

• Bayesian describe causal or apparently causal relationships

If (weight Δ =up & procedure=sham) then P(ALT=low)=0.6

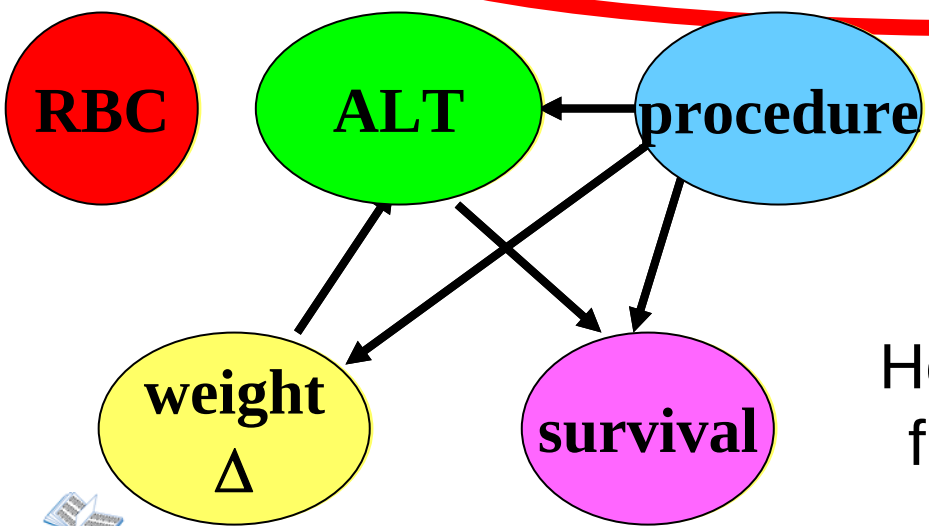


weight Δ	procedure	P(ALT) high	low
up	sham	0.4	0.6
down	sham	0.1	0.9
up	treatment	0.98	0.02
down	treatment	0.80	0.20
up	control	0.05	0.95
down	control	0.001	0.999

Bayesian Networks: General

• Bayesian networks are another way of expressing conditional probabilities

$$P(R,A,W,P,S) = P(R)\,P(A|\,W,P)\,P(P)\,P(W\,|\,P)\,P(S\,|\,A,P)$$



A model

How can we automatically find which model fits the data best?

Bayesian Networks: General

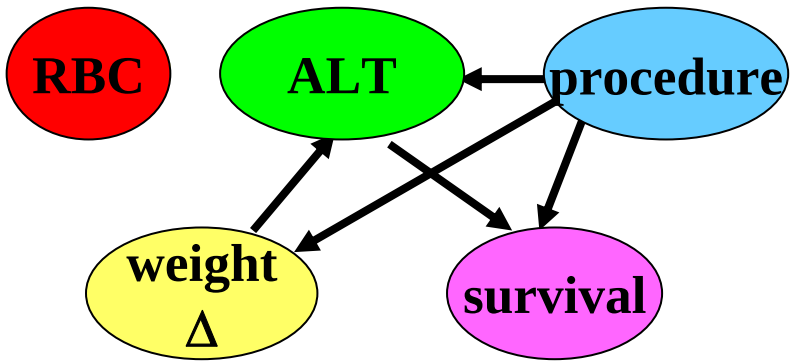
Model fit assessed with Bayes' Rule:

$$P(Model \mid Data) = P(Data \mid Model)\, P(Model) / P(Data)$$

P(Data | Model): evaluated using probability theory
P(Model): prior probability for each models or set of features
P(Data): probability of this particular data set given all other possibilities and knowing nothing of the model.

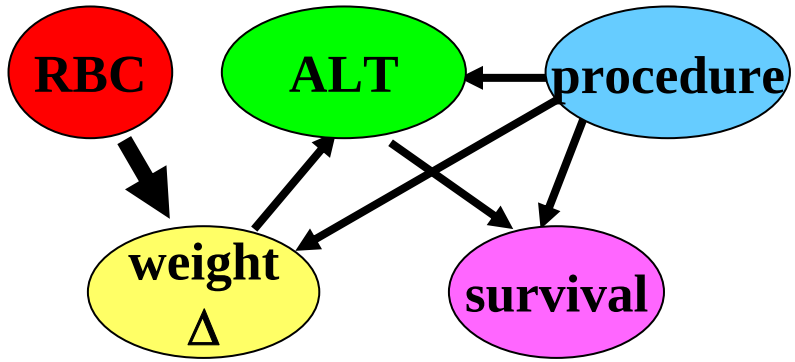
Goal: find the candidate structure that maximizes the Bayes score, $\log(P(Model \mid Data))$

Bayesian Networks: Searching



Model 1

P(Model 1 | Data)=0.0754



Model 2

P(Model 2 | Data)=0.0129

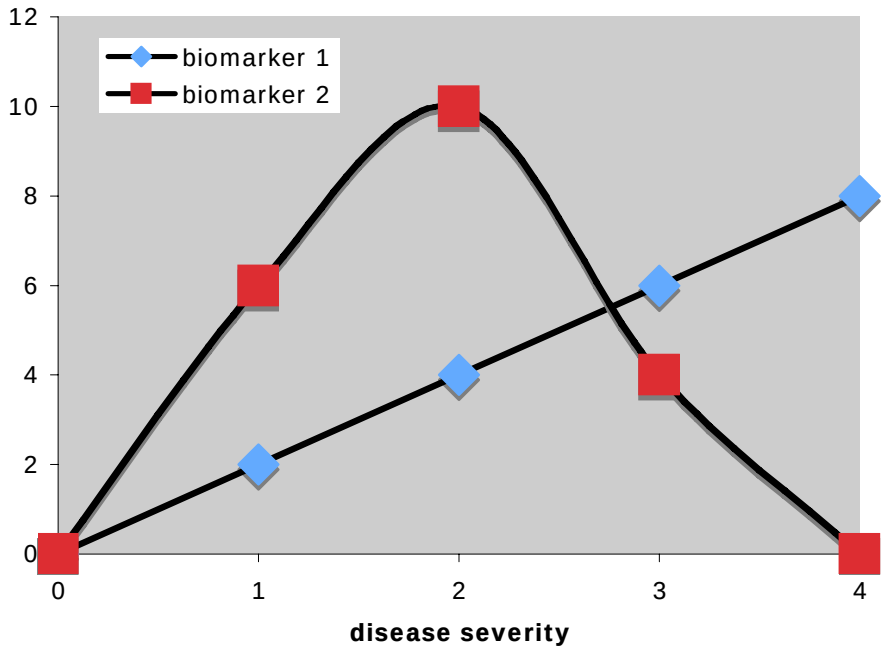
P(Model 1 | Data) > P(Model 2 | Data)
Model 1 fits the data better

Three Modeling Directions

1. Biomarker Identification
2. Static Bayesian Networks
3. Dynamic Bayesian Networks

Biomarker Identification

GOAL: Identify maximally relevant combinations of nonlinear predictors of any biological outcome.



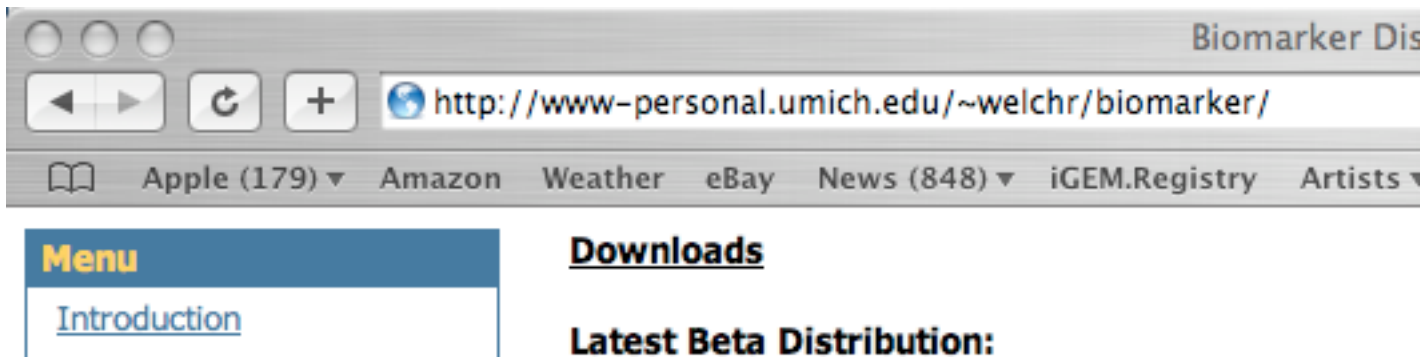
- Correlation methods identify linear relationships (#1), but can't find nonlinear relationships (#2)
- Many useful markers are likely nonlinear
- Short lists of biomarkers are often better

Biomarker Identification

- MARKIT uses a Bayesian network engine to identify biomarkers from hundreds of thousands of variables for 2 parents, or tens of thousands for 3 parents.

Beta Software available at:

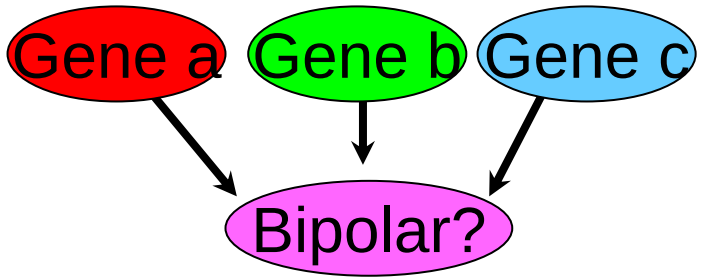
http://www-personal.umich.edu/~welchr/biomarker/



Biomarker Identification

Example 1: Molecular differential diagnosis of bipolar & schizophrenia

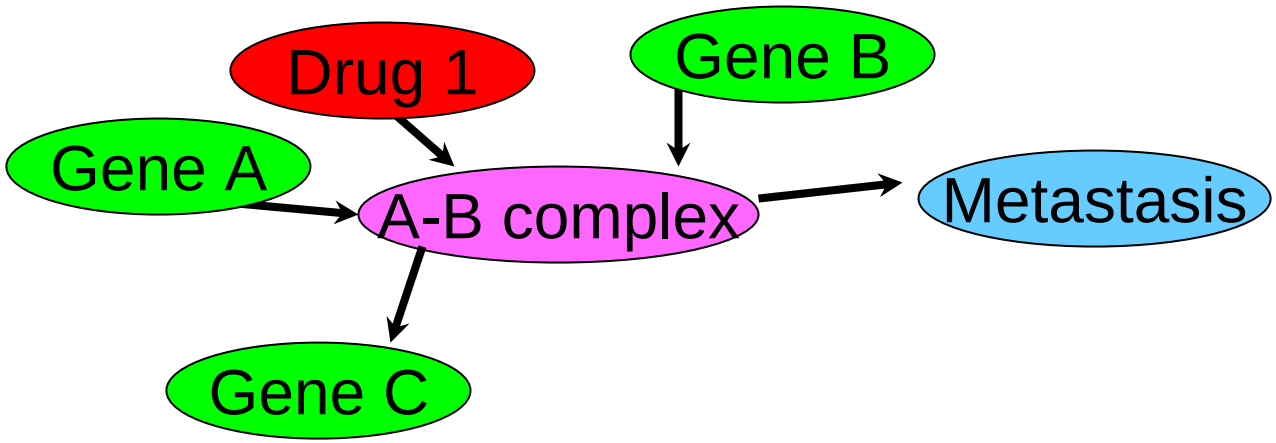
- 68 Affymetrix profiles of neutrophils from patients



Static Bayesian Networks

GOAL: Identify mechanistic pathways from data and user experience.

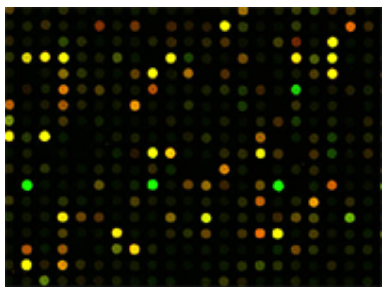
Focus area: Chinnaiyan data to identify plausible mechanisms in prostate cancer



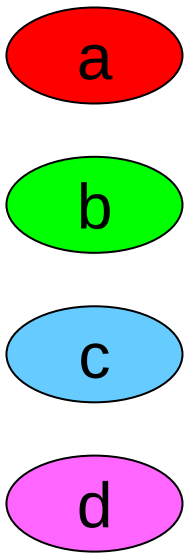
Data Integration: Promises and Pitfalls

- Most modeling efforts focus on a single data type

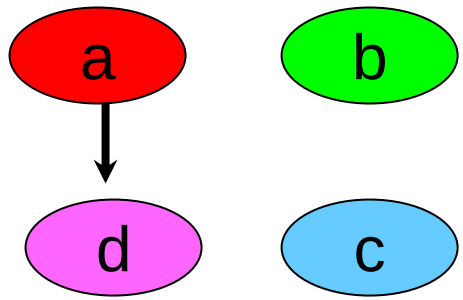
Data



mRNA

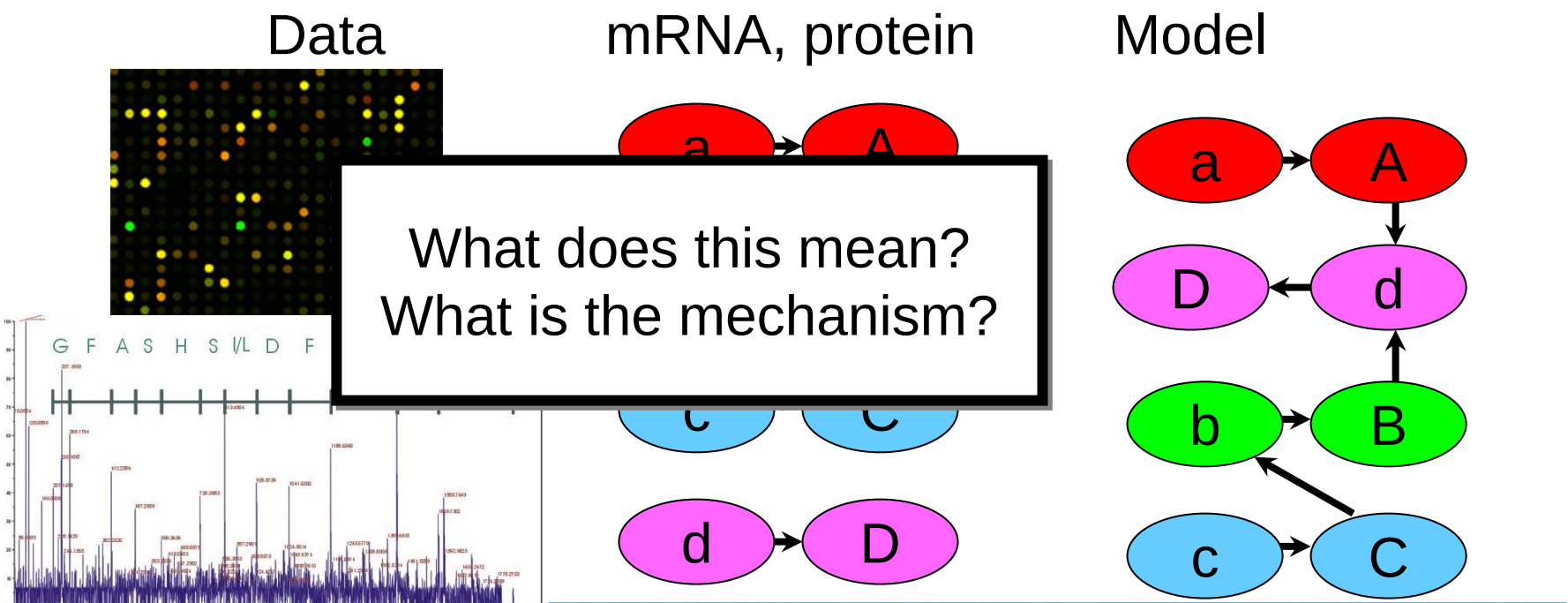


Model



Data Integration: Promises and Pitfalls

- Integrating more data makes the model more complete and complex



From http://www.med.monash.edu.au

Data Integration: Data Types

- Two kinds of are data possible:

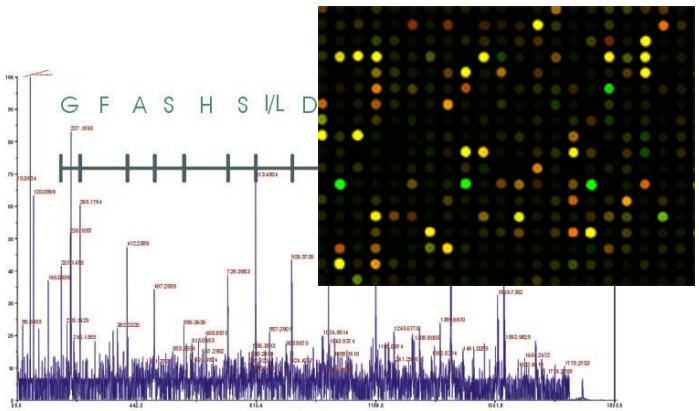
<u>Observational:</u>

Shows what **does** happen

?

<u>Relational:</u>

Constrains what **could** happen.



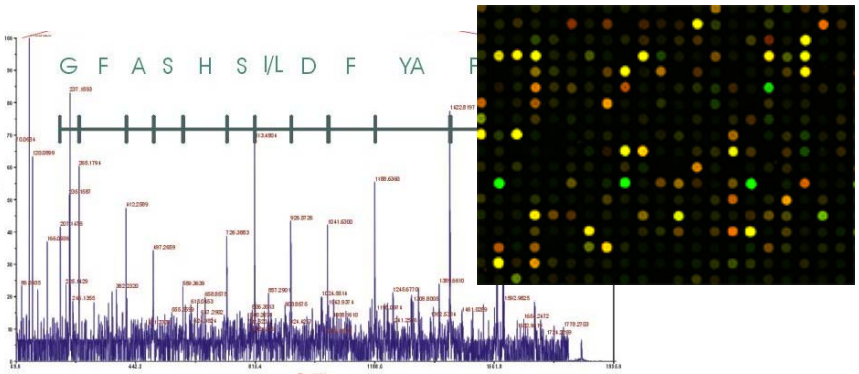
Data Integration:
Data Types

Relational data: Limits which mechanisms are possible.

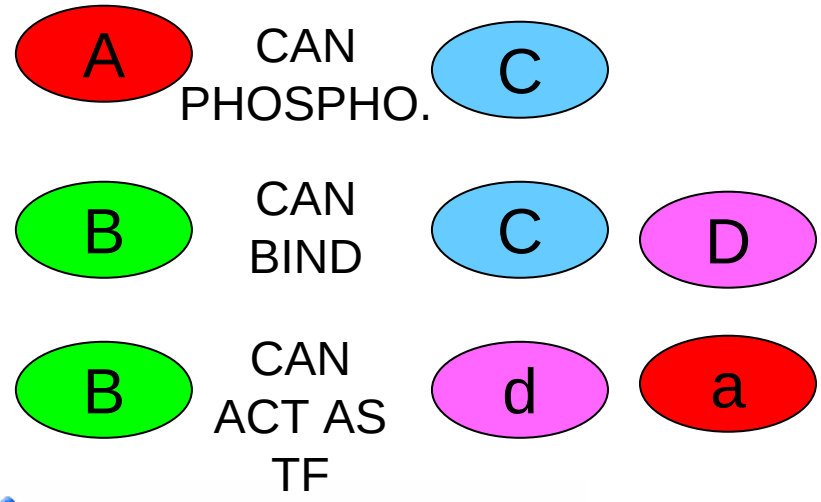
- Which proteins **can** form complexes?
- Which proteins **can** phosphorylate which targets?
- Which transcription factors **can** regulate which targets?

→ MiMI, NLP, GeneGo

Observational Data

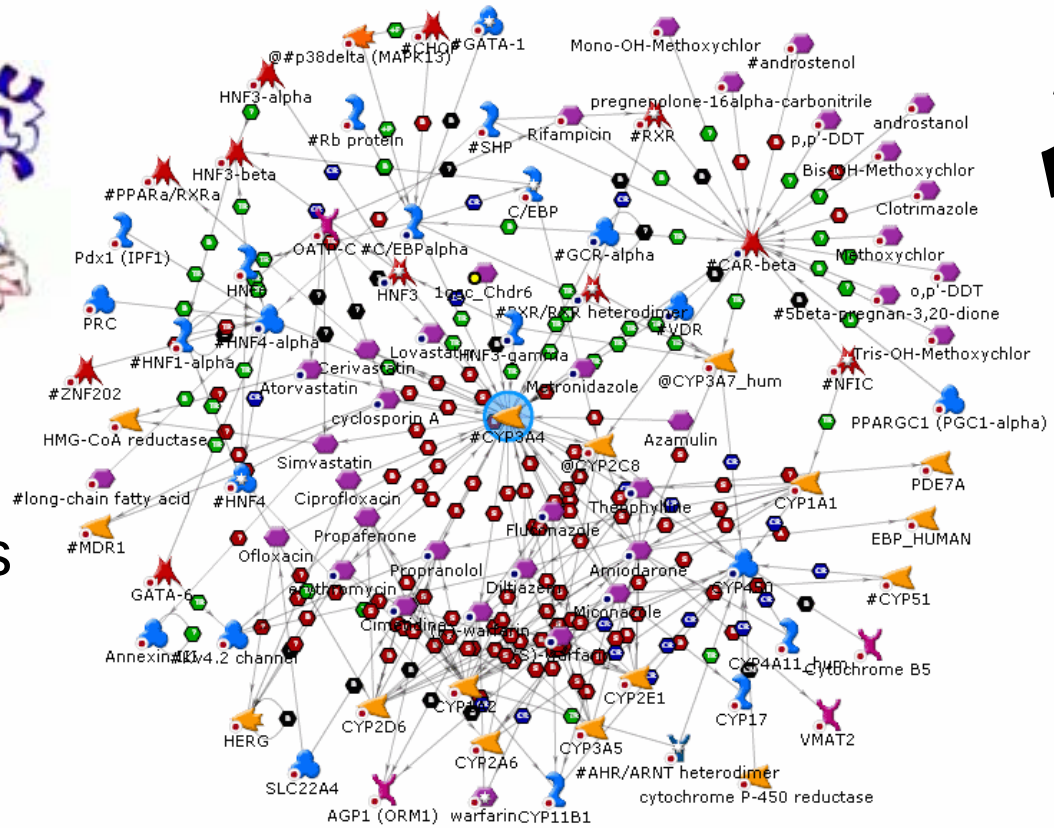


Relational Data



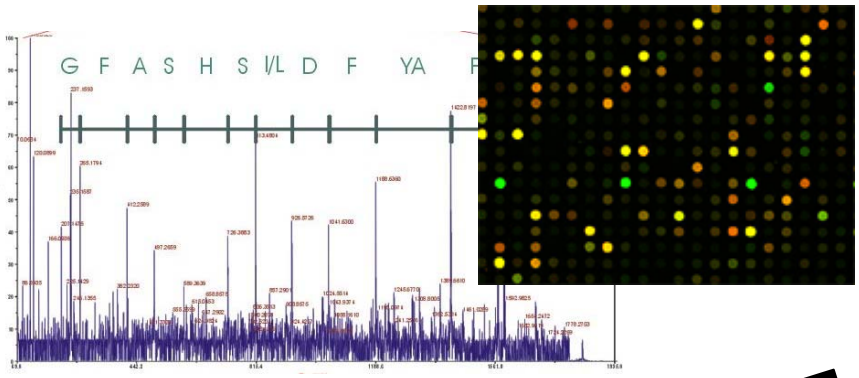
GeneGo

NLP resources

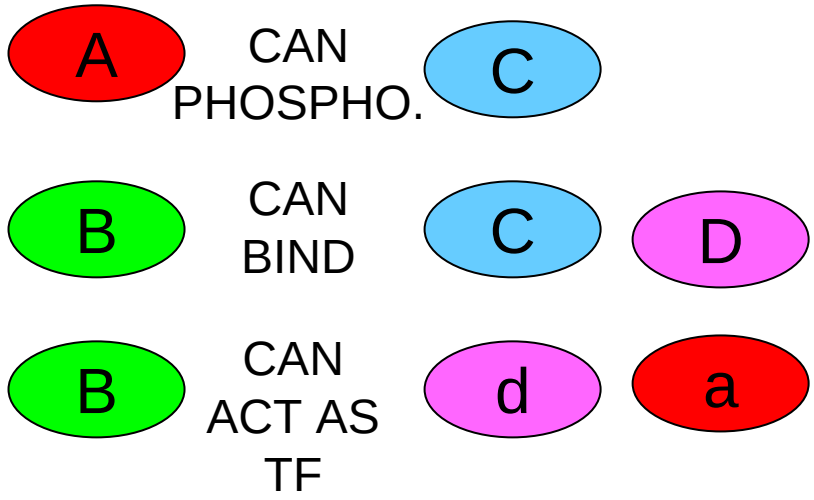


Ontology mapper

Observational Data

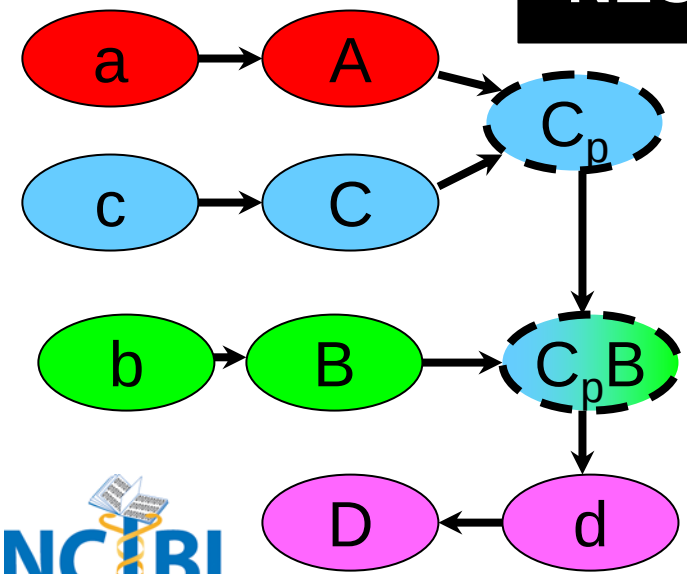


Relational Data



Possible Model:

NLC ?



Bubble

Java interface for learning, modifying, and interpreting Bayesian networks

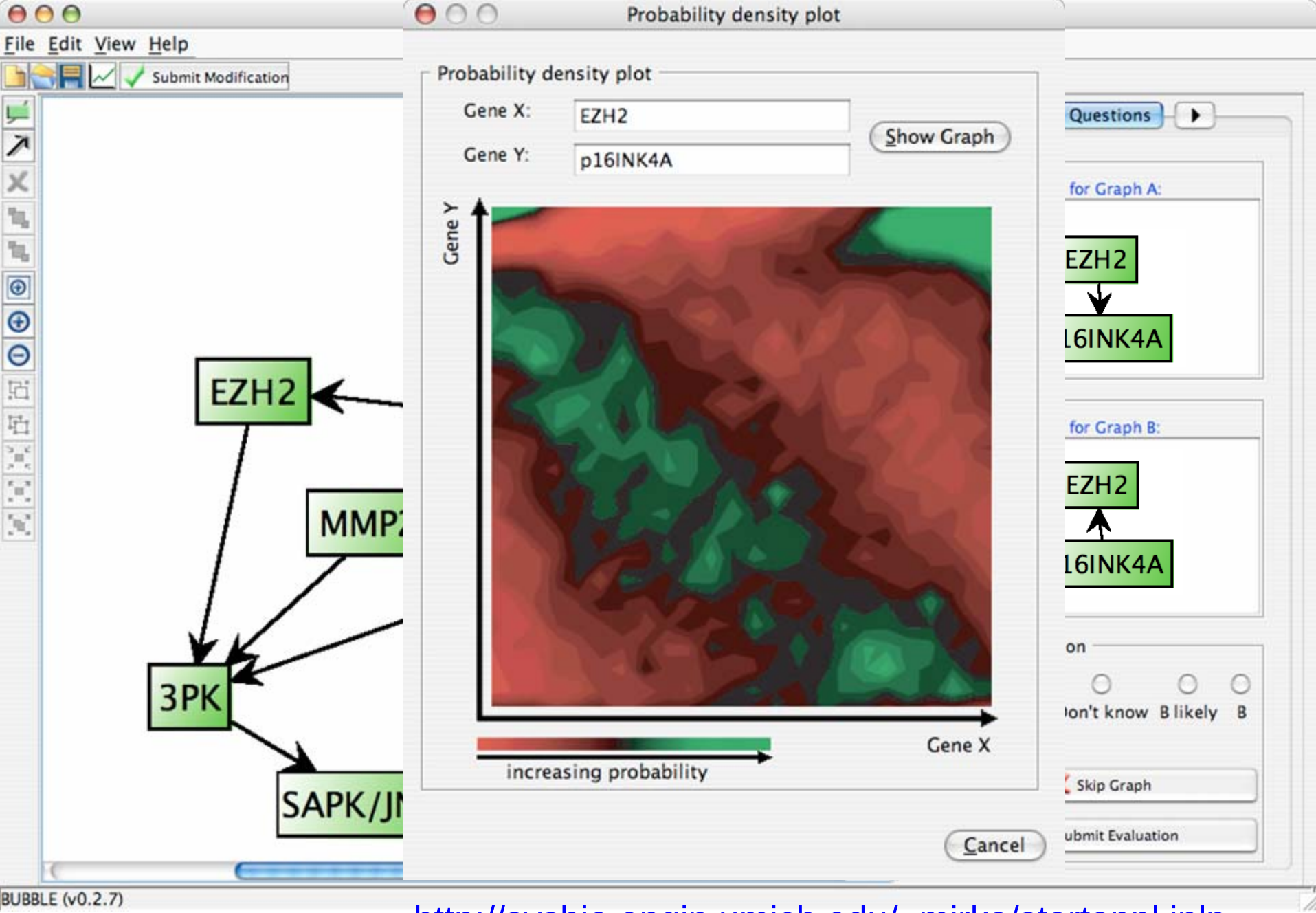
Beta version online:

http://sysbio.engin.umich.edu/~mirka/startappl.jnlp

OR

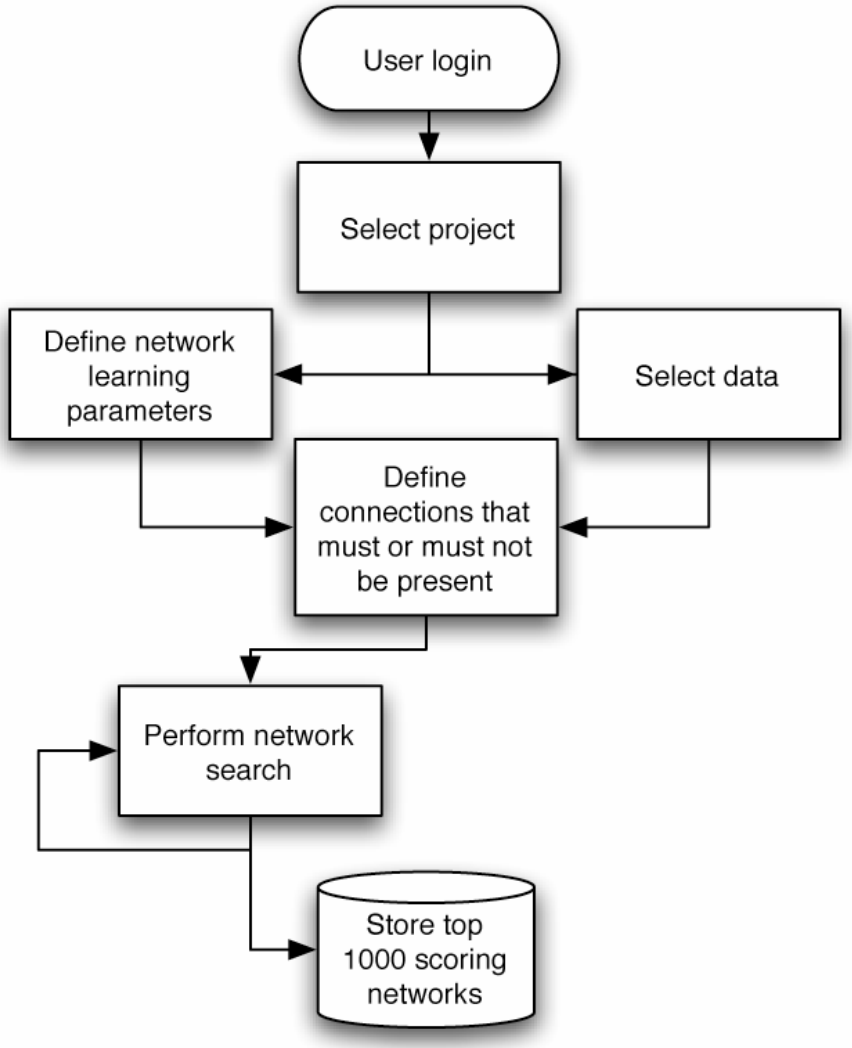
http://werewolf.engin.umich.edu:8080/

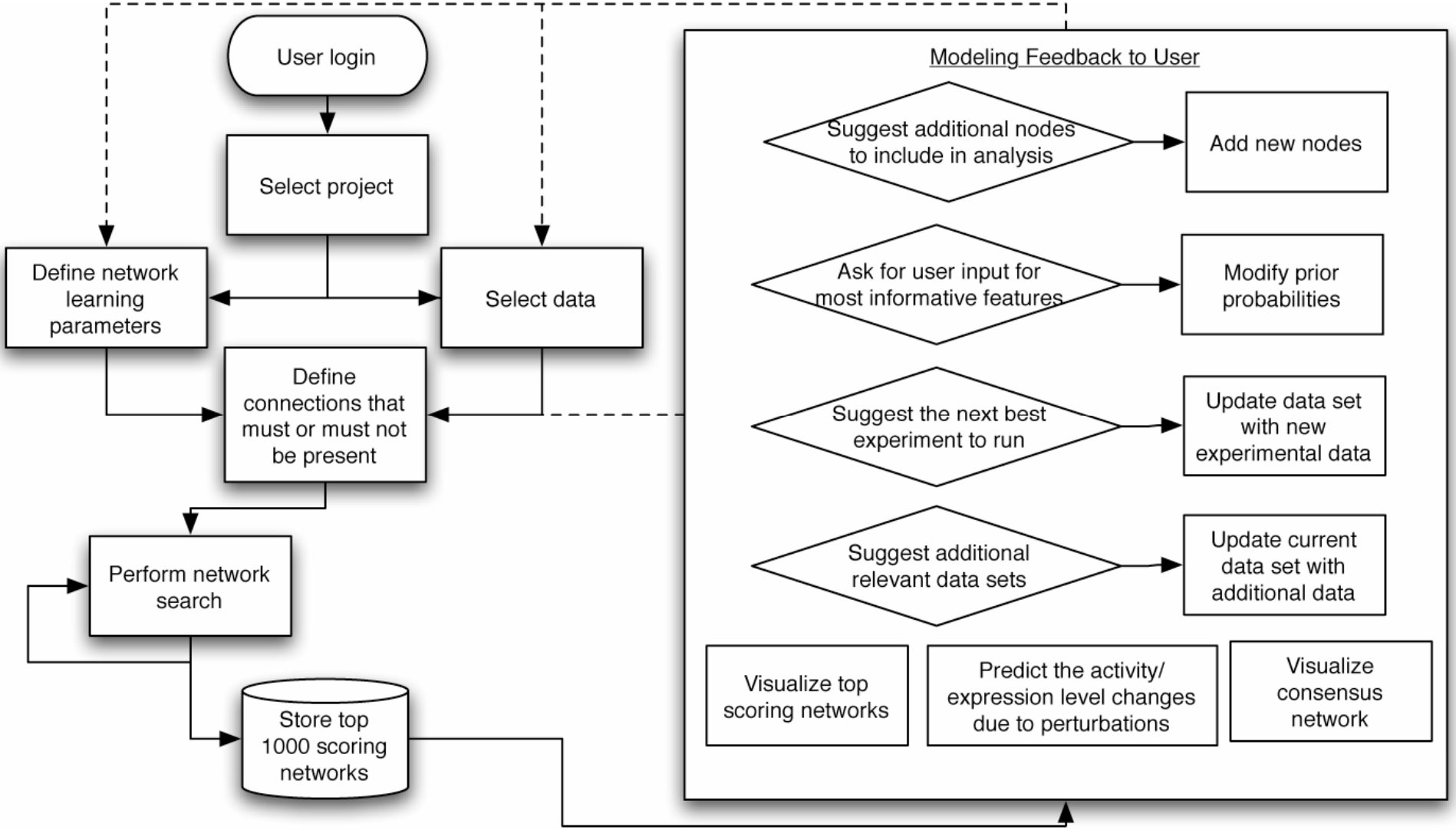
(less flexible, but more reliable)



http://sysbio.engin.umich.edu/~mirka/startappl.jnlp

http://werewolf.engin.umich.edu:8080/


User login
Select project
Define network learning parameters
Select data
Define connections that must or must not be present
Perform network search
Store top 1000 scoring networks


User login
Select project
Define network learning parameters
Select data
Define connections that must or must not be present
Perform network search
Store top 1000 scoring networks
Modeling Feedback to User
Suggest additional nodes to include in analysis
Add new nodes
Ask for user input for most informative features
Modify prior probabilities
Suggest the next best experiment to run
Update data set with new experimental data
Suggest additional relevant data sets
Update current data set with additional data
Visualize top scoring networks
Predict the activity/ expression level changes due to perturbations
Visualize consensus network

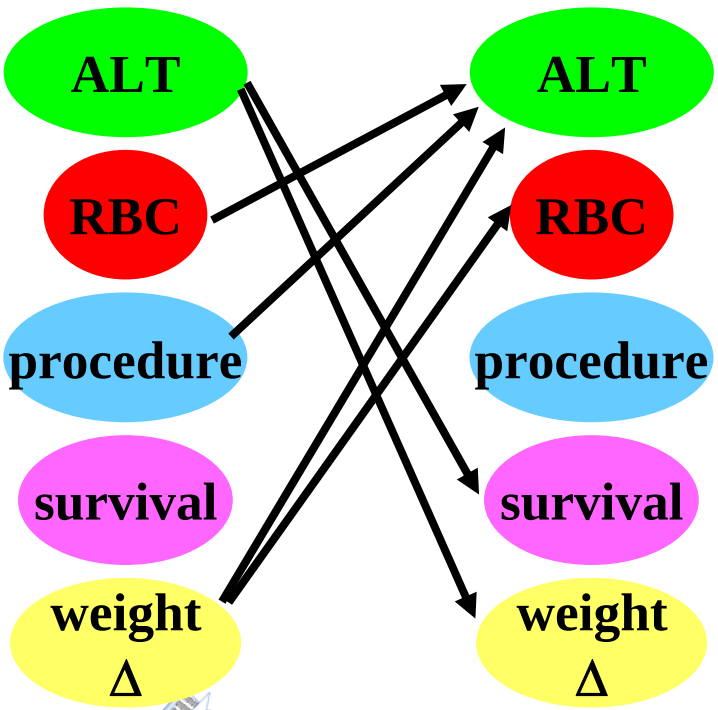
Dynamic Bayesian Networks

GOAL: Learn apparently causal mechanisms from temporal data (when we have it)

Dynamic Bayesian Networks

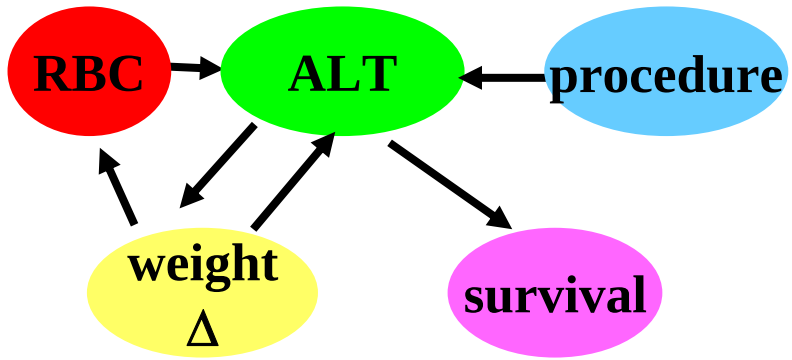
Unrolled Network

Yesterday (t_{i-1}) Today (t_i)

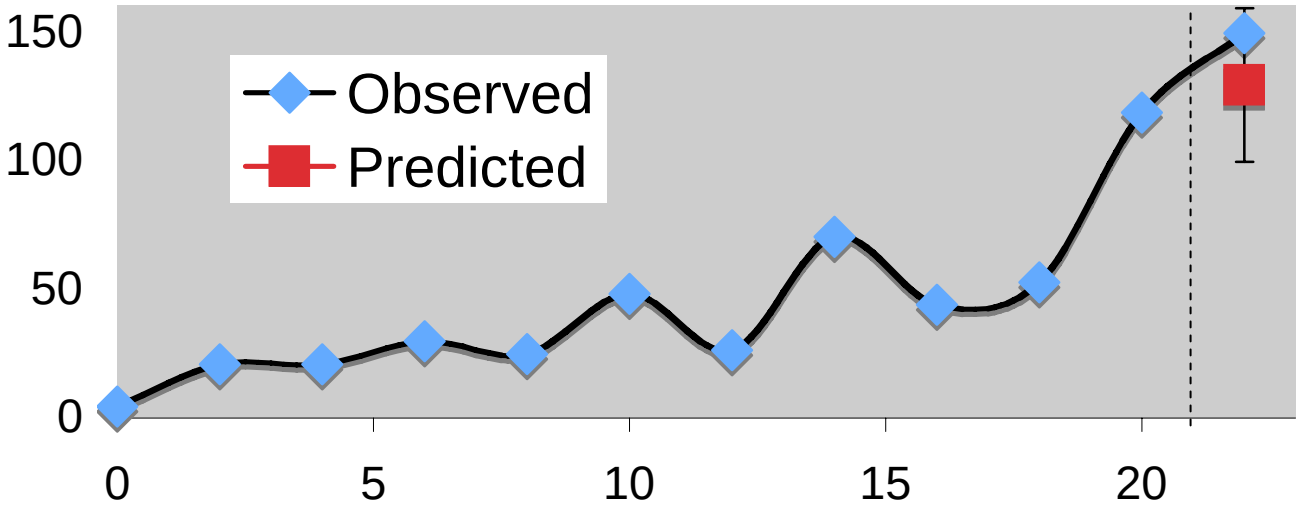


OR

Collapsed Network

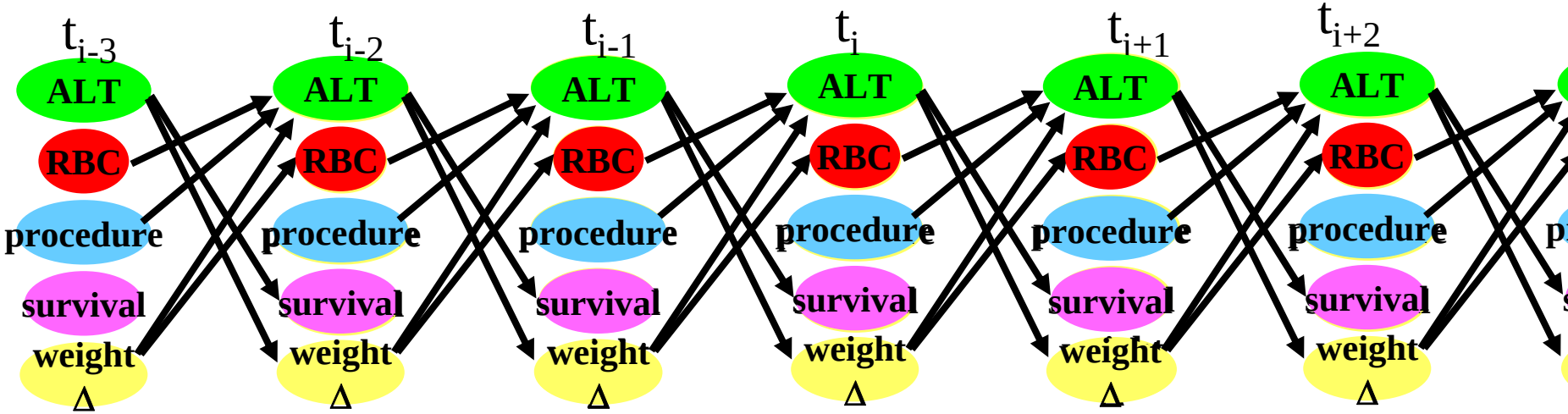


These are both examples of Dynamic Bayesian Networks (DBNs)

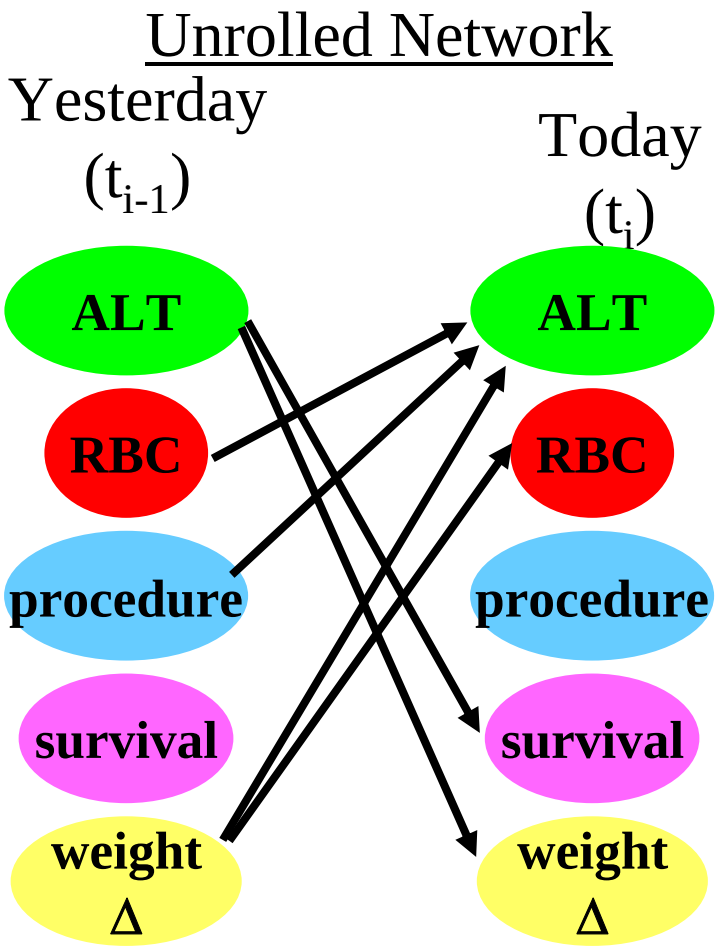


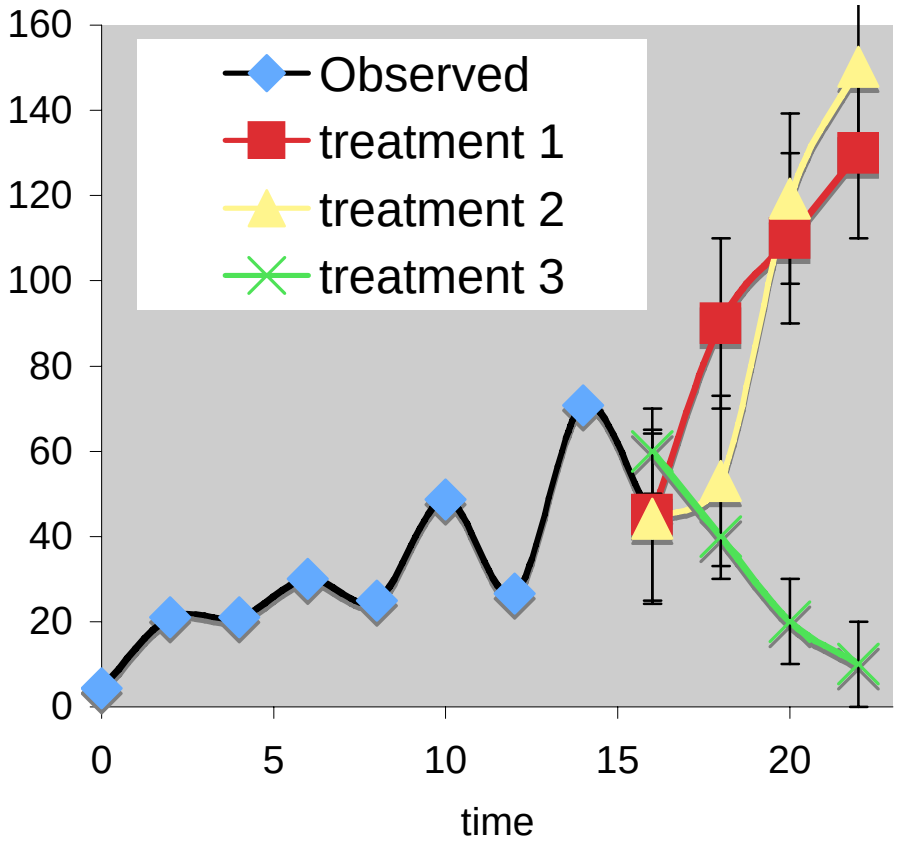
Predicts future responses

Model derived from past data



- Predictions can explore future alternatives

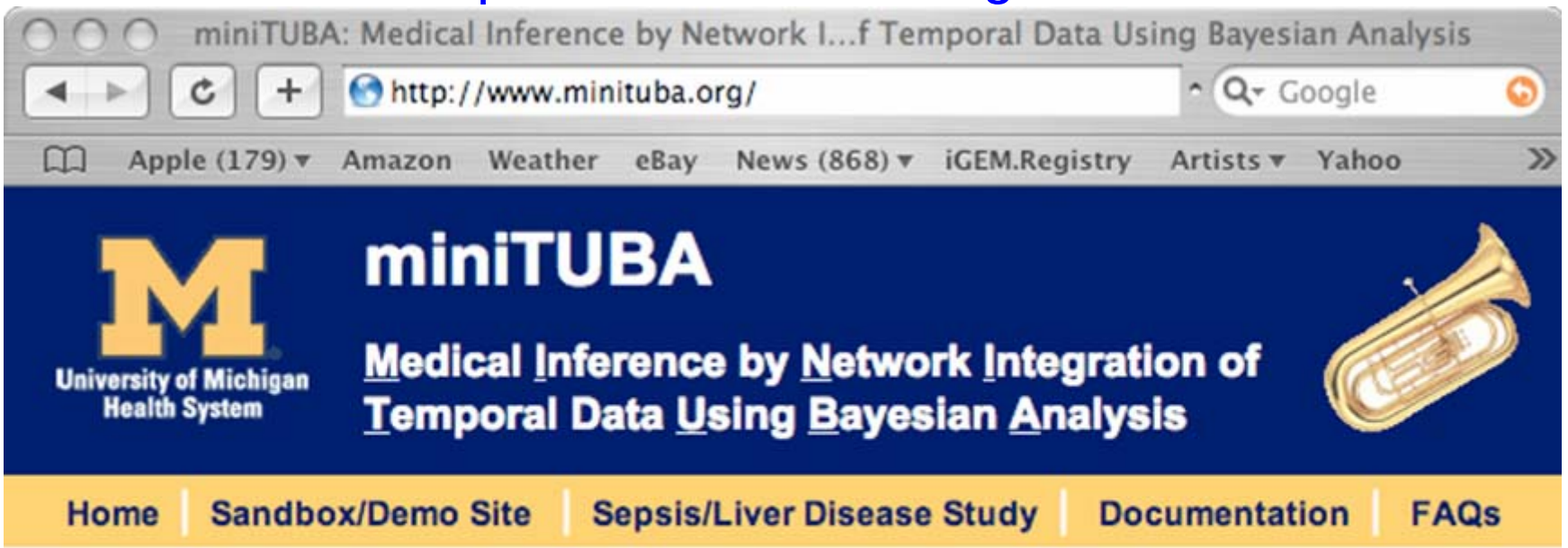




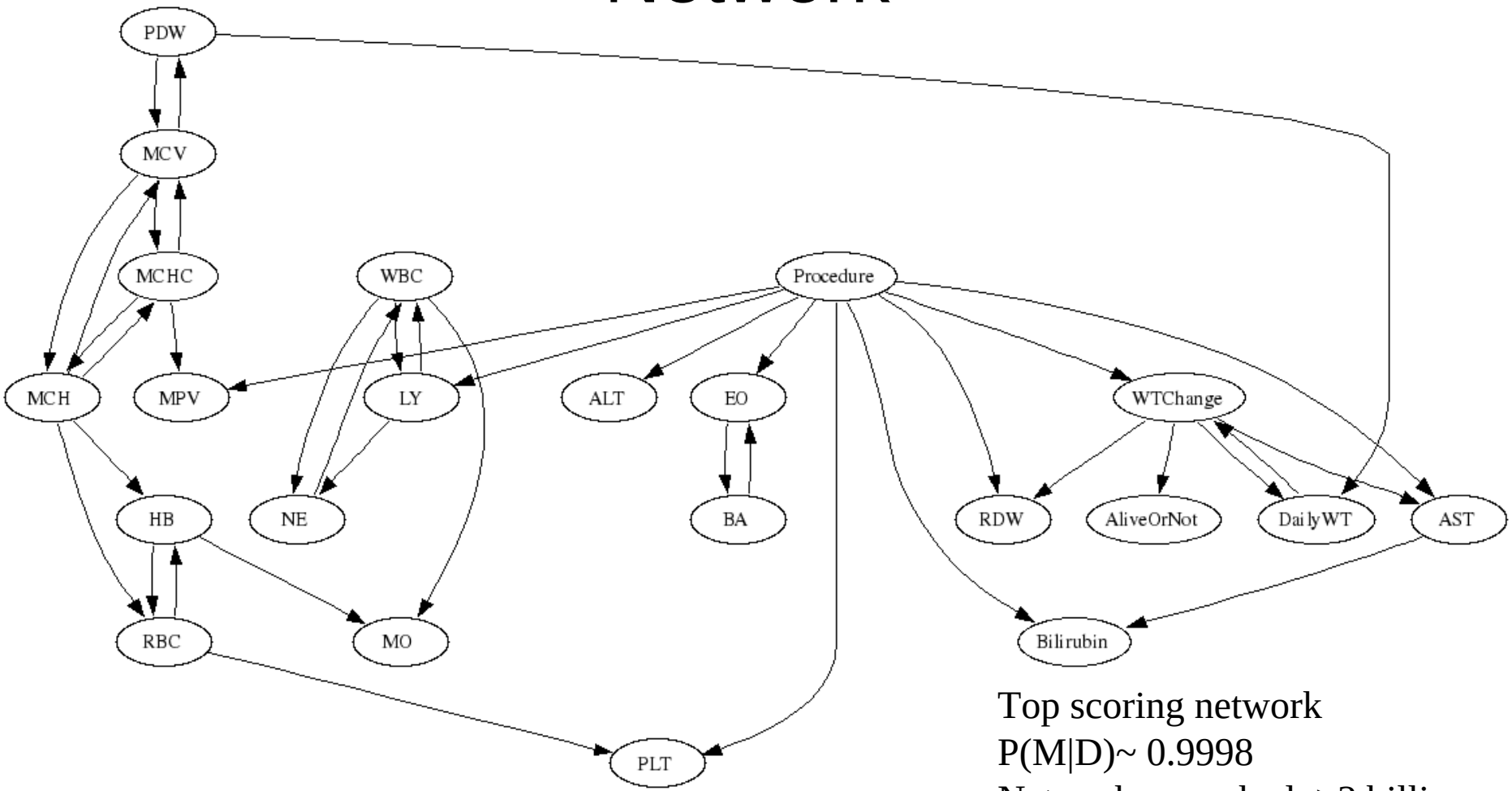
MiniTUBA

Web interface for learning and visualizing Dynamic Bayesian Networks

http://www.minituba.org/

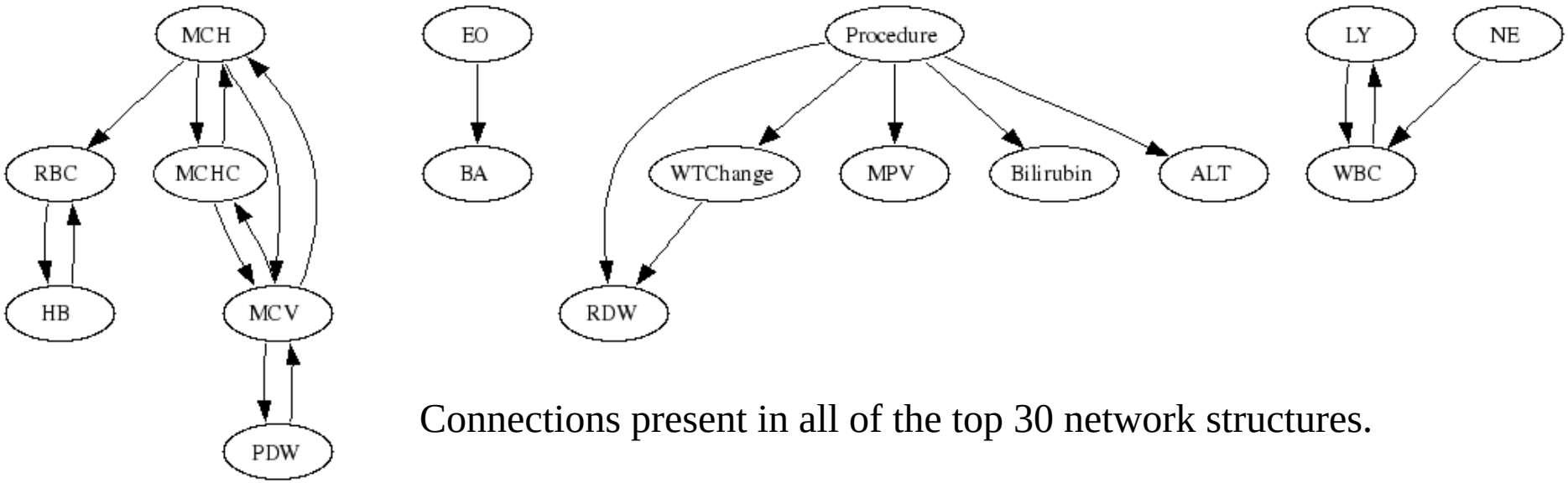


Top Scoring Dynamic Bayesian Network



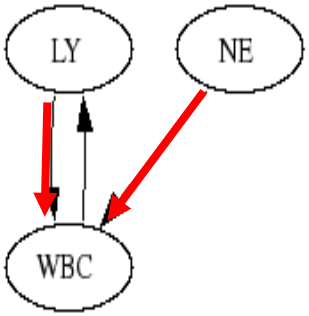
Top scoring network
P(M|D)~ 0.9998
Networks searched: >2 billion

Dynamic Bayesian Network Conserved Edges



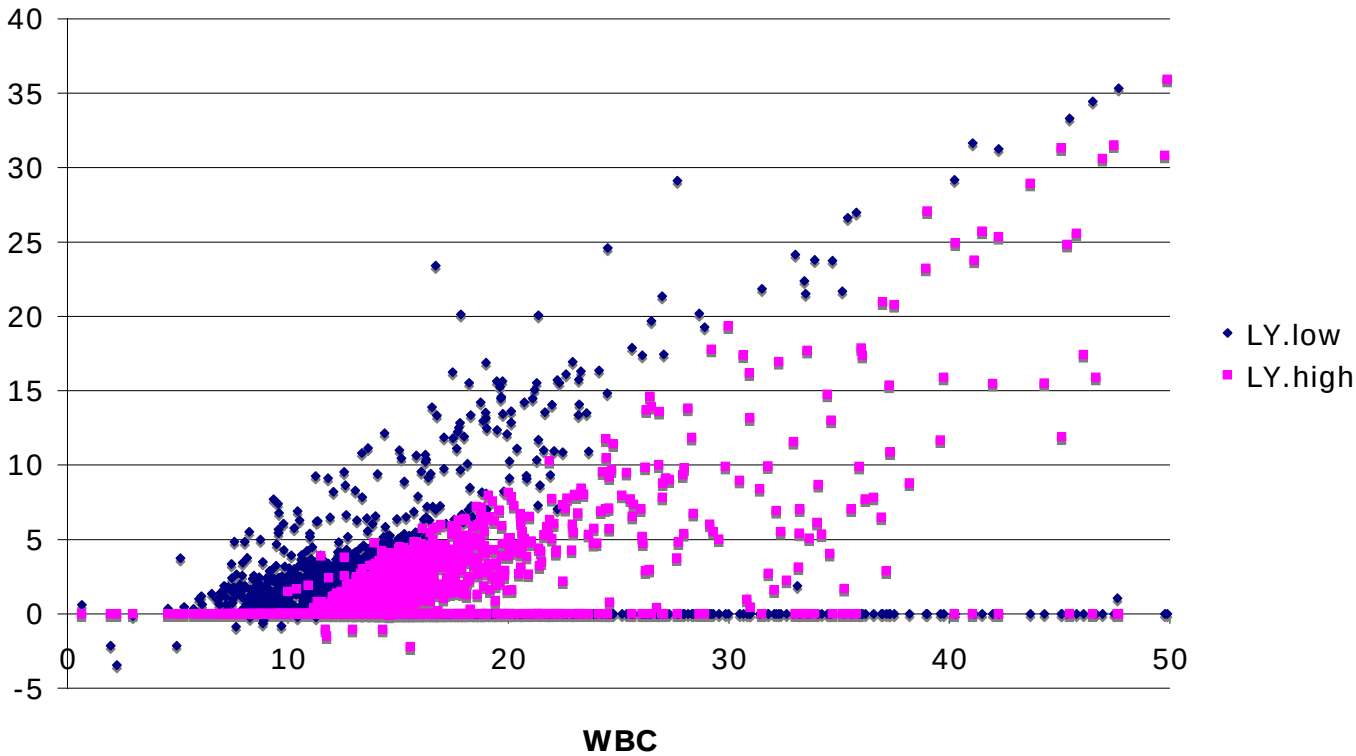
Connections present in all of the top 30 network structures.

Example Relationship: Lymphocytes (LY) and Neutrophils (NE) predict white blood cells (WBC)



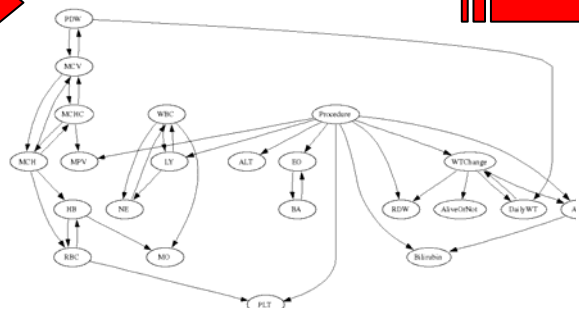
Approximate rule:

If(K*NE>WBC) then LY=low

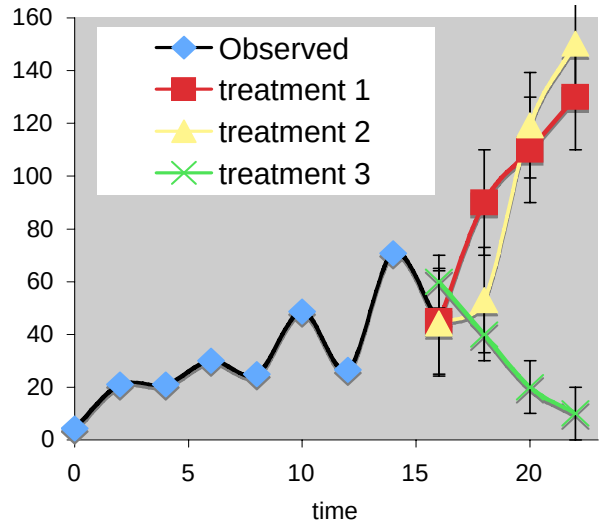


DBNs for clinical and experimental research: Future Vision

User enters data as they are gathered into a web based form



Data are incorporated into the existing dynamic Bayesian network



Model predicts the impact of various treatments

Model suggests treatment that maximizes survival probability or maximizes information content

Because model is constantly updating, it will dynamically learn and adapt based on all of the data

Questions?

To try the tools out go to:

Bubble:

http://sysbio.engin.umich.edu/~mirka/startappl.jnlp

http://werewolf.engin.umich.edu:8080/

miniTUBA:

http://www.minituba.org/

Bayesian Biomarker ID:

http://www-personal.umich.edu/~welchr/biomarker/

Or email me:

pwoolf@umich.edu


NCBI

Corrections:

• Find a way to set up for Matthias as the final punch

• Provide example from Bubble for Chinnayn data (how to use?)

• Color graph to show how this stuff works out.

Updates:

Start out with a vision both of how it is and how it should be. Then show how it works.

Biomarkers--fusion, and McInnins, and Feldman

Transcriptional networks--how do they work?

Bubble to DOOBN

•DBN

•Include email, focus on DBPs

•“Biology is an information science” by lee hood

• TF regulation of gene expression

•alanine transaminase=ALT test

•Ricardo Del Faveo in Colloumbia to study B-cell lymphoma. (working with Adndreas)

- •-lots of different geen expression data a

•Do DBPs really drive the computation, or is it something else?

•Odd that “biological discovery” and “biological modeling” are seen as different. Maybe this is my bioinformatics vs systems biology.

•Data are provided via wiki as a mechnanism.